单颗粒冷冻电镜三维重构研究的新进展

樊莉亚 邵书伟 王功明 万晓华 储琪 陈翔 张法

摘要：单颗粒冷冻电镜三维重构技术由于具有其他研究手段所不具备的优点，已经成为一种公认的研究生物大分子结构的强有力手段。然而，冷冻电镜显微图像的信噪比极低，要获得高分辨率的结果必须收集大量的图像数据，这使得冷冻电镜三维重构极其耗时。针对上述问题，本文详细介绍了单颗粒冷冻电镜三维重构的发展和现状，分析了当前所面临的主要挑战性问题，着重介绍了我们在单颗粒冷冻电镜三维重构相关研究工作上的进展。

关键词：冷冻电镜；三维重构；颗粒图像识别；自适应动态调度；ParaEMAN

1 引言

随着人类基因组和一系列模式生物全基因组计划的完成，生物学家发现想要更好地理解生命过程，就必须对基因的表达产物蛋白质进行结构和功能上的研究。按照"结构决定功能"的基本原理，只有清楚了以蛋白质为主体的生物大分子的三维结构，才有可能最终确定它们的生物功能。因此以 X 射线晶体学、核磁共振技术、电子显微学和计算生物学为基本研究手段的结构生物学将扮演越来越重要的角色[1]。

电子显微学是通过电子显微镜技术并结合图像处理技术发展起来的。同 X 射线晶体学和核磁共振技术这两种成熟的结构生物学研究手段相比，电子显微学具有可直接获得分子的形貌信息和相位信息，能够解析那些不适合应用 X 射线晶体学和核磁共振技术进行分析的蛋白质等优点。随着生物样品制备技术的完善，电子显微镜设备的进步以及数字图像处理技术的发展，电子显微学已经成为一种公认的研究生物大分子、超分子复合体及亚细胞结构的有力手段[2]。

电子显微学方法又包含三种独立的技术：电子晶体学、单颗粒三维重构和电子断层成像技术。其中单颗粒三维重构技术，由于只需要处理同一大分子随机散布的电镜照片，没有形成晶体的要求，且分子越大重构结构越好，已经成为结构生物学研究中至关重要和不可替代的研究手段。目前单颗粒三维重构技术解析的最好分辨率已达到 4 埃（Å）左右[3]。但是冷冻电镜显微图像的信噪比极低，要获得高分辨率的结果必须收集大量的图像数据（近 10 万个颗粒的图像）。此外，每一张冷冻电镜的颗粒图像都需要计算其投影方向和进行快速傅立叶变换等多种处理，总的重构时间极其耗时，通常需要约 10^6CPU 小时，而且目前只能处理 20,000~40,000 张颗粒图像，详见文献[4]。因此借助高性能计算系统开发快速准确的三维重构方法就显得尤为迫切和关键。

为了加强冷冻电子显微技术的相关应用和研究，中国科学院蛋白质科学研究平台生物成像中心即将安装世界上最先进的低温场透射电子显微镜 Titan Krios，同时 2008 年启动的中国科学院知识创新工程重大项目《面向蛋白质科学的高性能计算研究》也将单颗粒冷冻电镜三维重构作为其两大应用之一。

2 单颗粒冷冻电镜三维重构

电镜三维重构的思想早在 1968 年就由德罗西耶（D.De Rosier）和克卢格（A.Klug）[5]提出，而冷冻电镜技术是在 1974 年首次由泰勒（K. Taylor ）和格莱瑟（R.M.Glaeser ）创

建[6]，经过三十多年的发展，冷冻电镜技术已经成为研究生物大分子结构与功能的强有力手段。冷冻电镜三维重构技术主要是将相同的生物大分子样品保存在液氮或液氦温度下，利用透射电子显微镜进行二维成像，再经过对二维投影图像的分析进行三维重构[7]。

2.1 电镜三维重构原理

德罗西耶和克卢格提出的三维重构理论是借助一系列沿不同方向投影的电子显微像来重构被测物体的立体构型，他们提出了利用数字图像处理技术进行电子显微像三维重构测定生物大分子结构的概念和方法。电镜三维重构思想的数学基础是中央截面定理和傅立叶变换。中央截面定理的含义是：一个函数沿某方向投影函数的傅立叶变换等于此函数的傅立叶变换通过原点且垂直于此投影方向的截面函数[2]。因此电镜三维重构的理论基础是一个物体的三维投影像的傅立叶变换等于该物体三维傅立叶变换中与该投影方向垂直的，通过原点的截面（中央截面），如图 1 所示[8]。每一幅电子显微像是物体的二维投影像，沿不同投影方向拍摄一系列电子显微像，经傅立叶变换会得到一系列不同取向的截面。当截面足够多时，会得到傅立叶空间的三维信息，再经傅立叶反变换便能得到物体的三维结构，如图 2 所示[2]。这种方法目前已经在很广泛的范围内得到应用，从无固定结构特征的细胞器[1]和生物大分子复合物到大分子晶体，已发展为蛋白质结构解析的一种实用方法。

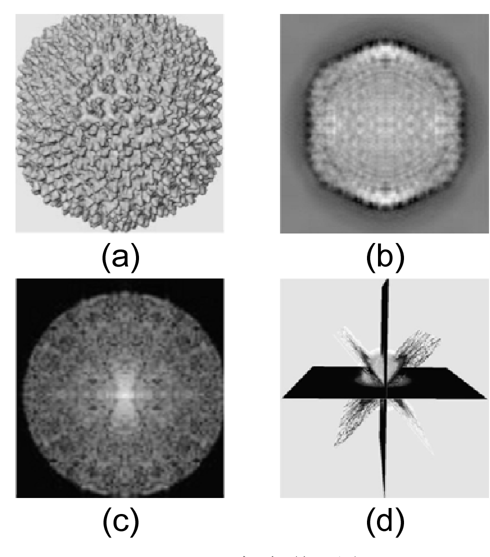


图1. 中央截面定理

图中**(a).** 生物大分子的三维模型; **(b).** (a)中的三维模型在某一方向的投影; **(c).** 投影的傅立叶变换; **(d).** 三维模型二维投影的傅立叶变换等同于原三维模型傅立叶变换的中央截面，图中的 4 个截面分别对应于 4 个不同方向的投影(0°,~45°,90°,~135°)

2.2 单颗粒冷冻电镜三维重构原理

单颗粒冷冻电镜技术是获得生物大分子三维重构图像的重要方法。所谓单颗粒法就是对分离纯化后的颗粒状分子进行结构分析。其基本原理是：通过对相同的生物大分子某方向的投影显微像在实空间中经过调整后进行叠加平均，从而提高信噪比，使粒子中共同部分的结构信息得到加强,最后对各种不同投影方向的单颗粒显微像在三维空间中进行重构，从而获得单颗粒大分子的三维结构信息。单颗粒冷冻电镜的主要步骤如图 3 所示[9]：（1）. 制备化学和结构上均一的生物大分子的冰冻含水样品；（2）. 选择最有可能产生最佳图像的最佳颗粒密度和玻璃态冰厚度的样品；（3）. 设定最佳的参数（比如：欠焦值、放大倍数和电子剂量等），拍摄并记录这些样品区域的大量图像；（4）. 用手工或半自动程序选择离散分子形成的投影图；（5）. 通过各种图像处理的方法计算不同图像之间的相对方位，进而重构出生物大分子的三维结构模型；（6）. 最后结构分析和评价，将从晶体学或核磁共振获得的蛋白质结构的原子坐标定位到三维结构密度图中。

图 4 概括了单颗粒冷冻电镜三维重构技术涉及到的从二维投影图像到三维重构模型的

[1] 细胞质中具有一定结构和功能的微结构，如： 线粒体；叶绿体；内质网；高尔基体；核糖体；溶酶体；液泡；中心体等。

复杂的图像处理过程[8]：

图像获取：由于电子显微镜照片的信噪比非常低，要获得高分辨率的分子三维模型，必须采集尽可能多的电子显微颗粒图像。当前主要的方法是人工挑选，这是一项耗时乏味的体力活，若用人工挑选几十万张颗粒图像几乎是不可能的事情。此外，该阶段还需要对图像进行降噪处理和欠采样处理，详细资料请参见[10]。

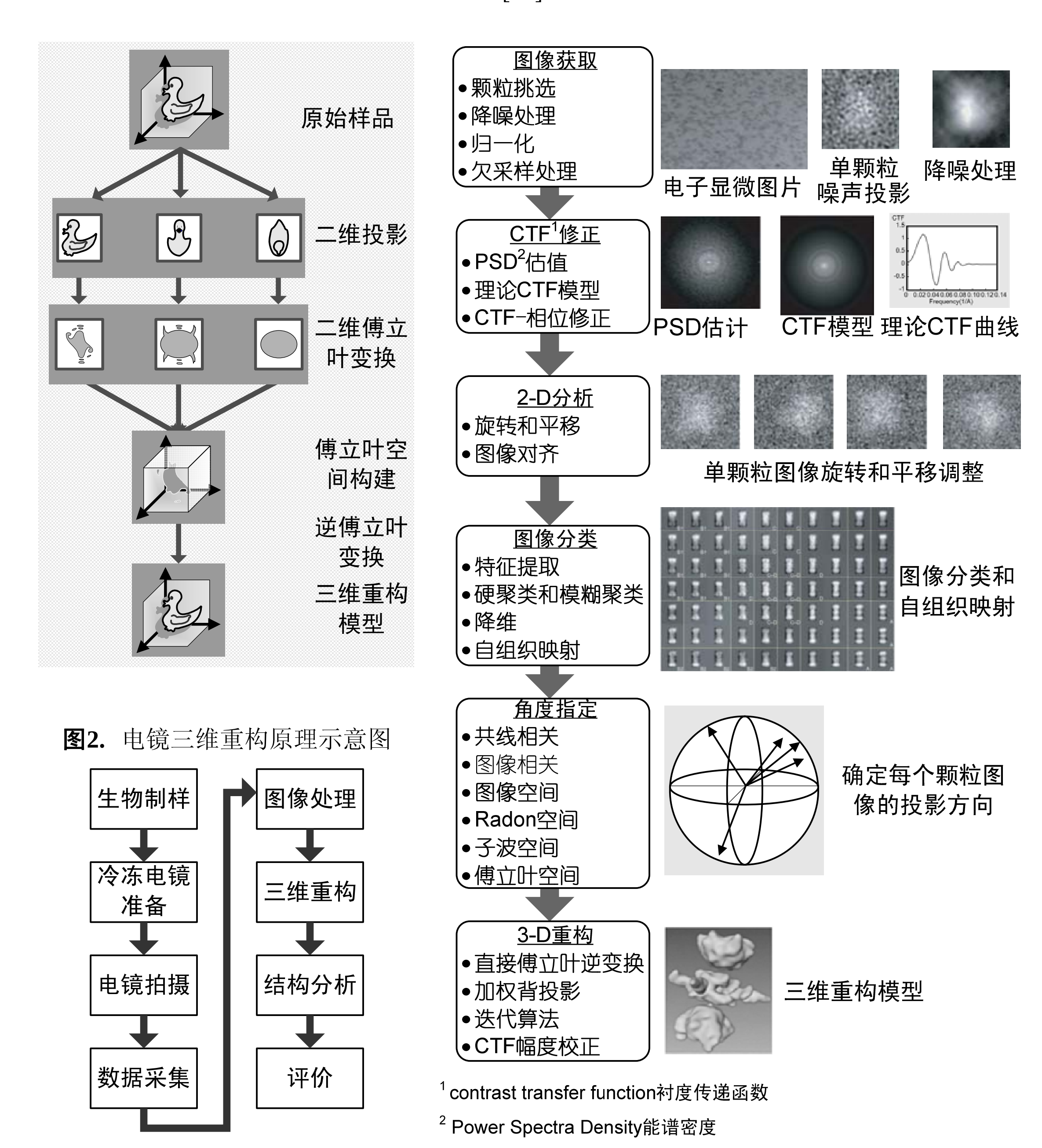


图2. 电镜三维重构原理示意图

图3. 单颗粒冷冻电镜的操作步骤

图4. 单颗粒冷冻电镜三维重构的处理流程

衬度传递函数修正：透射电镜的物镜不是理想的凸透镜，由于球差、离焦照相等因素的影响，最终的显微图像是经过衬度传递函数（contrast transfer function, CTF）和一些噪声函数作用的结果，并不是真正的样品投影势。因此重构时必须对图像进行 CTF 修正，常用的方法是首先估算颗粒图像的功率谱密度(power spectral density, PSD)，然后利用理论的 CTF 模型对 PSD 进行拟合，进而修正 CTF-相位[11,12]。

图像二维分析：在获得颗粒图像数据集的基础上，需要对每一个图像进行旋转、平移和对齐(alignment)等处理，以得到颗粒图像每个方向的平均图、对称性等信息。

图像分类：在进行三维重构处理之前，必须对颗粒投影图像进行分类，以保证每一类中所有的图像都属于同一方向的投影图，否则将会对重构结果产生非常不利的影响[13]。通常是采用经典的模式识别和聚类技术，例如特征提取、自相关/互相关分析、硬聚类和模糊聚类等等[2,14,15]。

角度指定：颗粒图像分类完成后，需要计算每一类图像的投影方向。一般是通过比较颗粒图像和计算机模拟生成的投影来确定颗粒图像的投影方向。为了降低噪声的影响，通常用每类颗粒图像的平均图来代表该类。

三维重构：根据中央截面定理，每个颗粒图像的傅立叶变换等同于原模型三维傅立叶空间中一个中央截面。因此根据每类颗粒图像的投影方向可以重构出三维傅立叶空间，然后采用直接逆傅立叶变换或者加权背投影（weighted back-projection）经过多次迭代重构优化，就可以最终获得分子的三维结构模型。

2.3 单颗粒冷冻电镜三维重构软件 EMAN

EMAN 是由美国国家大分子图像中心的路德克（Steven J. Ludtke）等人开发，于 1999 年推出第一个版本[16]，如今已成为世界上使用最广泛，结果分辨率最高的单颗粒重构软件之一。使用 EMAN 对电子显微镜照片进行单颗粒三维重构包括以下三个基本步骤：（1）．颗粒挑选，从电子显微镜照片中挑选出生物样品颗粒的图片，并将所有挑选的结果进行保存；（2）．初始模型生成，利用上一步得到的颗粒图片，生成一个初始三维模型。该模型一般较为粗糙，通常无法满足预期的分辨率要求；（3）．模型优化，通过迭代的方式对初始模型逐步优化，直到三维模型的分辨率满足要求或者迭代收敛为止。以上步骤中，模型优化是最为耗时的，它占到单颗粒重构总时间的绝大部分，同时它直接决定了最终分子结构的分辨率。

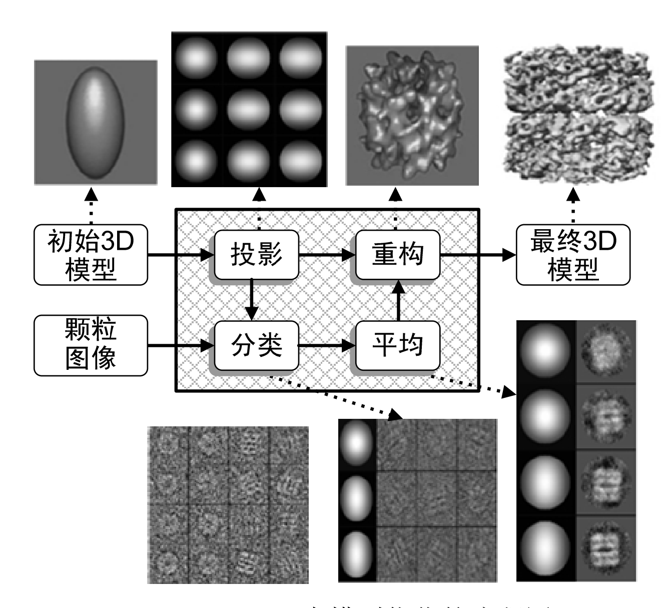


图5. EMAN 中模型优化的流程图

模型优化的每一轮迭代需要经历四个主要操作，如图 5 所示[17]：

第一步**投影操作**：输入是一个三维模型。对于第一轮迭代，该模型是上一步生成的初始模型；对其它轮迭代，该模型来自上一轮迭代生成的新模型。投影操作的目的是计算输入模型沿着一组欧拉角$\left(\alpha_i,\beta_i,\gamma_i \quad i=1,2,\cdots m\right)$的二维投影。这些投影将作为模板，在后续操作中与分子颗粒图像进行比对。在后面的讨论中，假设该操作生成的二维投影为$T_1,T_2,\cdots T_m$，其中T_i对应欧拉角$\left(\alpha_i,\beta_i,\gamma_i\right)$。

第二步**分类操作**：评估各分子颗粒图像$\left(P_1,P_2,\cdots P_n\right)$与上一步生成的各投影之间的相似度，并以此为依据，对分子颗粒图像进行分类。对于颗粒图像P_j $\left(1\le j\le n\right)$，具体算法如下：

给定任意投影 T_i，首先将 P_j 与 T_i 进行旋转平移对齐，找到最佳的旋转角度和平移向量，使得 P_j 经过该角度和向量的旋转平移之后，与 T_i 最为相似。之后将 P_j 和 T_i 对应像素点的灰度值分别作为二维平面上点的横坐标与纵坐标，对这些二维平面上的点进行最小二乘直线拟合。T_i 与 P_j 的相似度 $s(i,j)$ 由拟合误差决定，拟合误差越大，相似度越低，反之亦然。最后找出与 P_j 相似度最高的投影 $Tc(j)$，即满足如下条件的投影：

$$s(c(j),j)=max\{s(i,j)|1\le i\le m\}$$

将 P_j 归入第 $c(j)$ 个颗粒图像类，并把 $(\alpha_{c(j)},\beta_{c(j)},\gamma_{c(j)})$ 作为 P_j 的投影取向。对于每一个颗粒图像重复上述过程将得到一组颗粒图像类 $C_1,C_2,\cdots C_m$，它们包含的颗粒图像满足条件：

$$C_i=\{P_j\,|c(j)=i\}\quad i=1,2,\cdots m$$

第三步**平均操作**：对以上生成的每一类颗粒图像分别进行，输出每一类颗粒图像的平均图。具体的算法分两步：

（1）筛选颗粒图像。通过计算阈值 $cut(i)$，将相似度低于该值的颗粒图像从类中删除。因此，经过筛选之后的类 C_i’包含的颗粒图像满足条件：

$$C_i^{'}=\{P_j\,|c(j)=i\wedge s(i,j)>cut(i)\}\quad i=1,2,\cdots m$$

（2）利用筛选之后剩余的颗粒图像计算该类的平均图。平均图每一像素点的灰度值是经过筛选之后的类中所有颗粒在该点处灰度的均值。假设由颗粒图像类 C_i 生成的平均图为 a_i，由于 C_i 对应的投影 T_i 以及 C_i 中颗粒图像的投影角度均为 $(\alpha_i,\beta_i,\gamma_i)$，因此认为 a_i 的投影角度也为 $(\alpha_i,\beta_i,\gamma_i)$。

第四步**重构操作**：利用上一步得到的所有平均图生成新的三维模型。首先计算各类平均图的傅立叶变换，得到一组频谱图。之后将各频谱图插入到三维傅立叶空间中。根据中央截面定理，每一个频谱图对应三维傅立叶空间中一个经过原点的截面，并且该截面与平均图对应的投影方向垂直。当所有频谱图都按照正确的位置插入到三维傅立叶空间中后，对傅立叶空间中的三维体数据进行傅立叶逆变换，即可得到新的三维模型。

2.4 当前研究面临的问题

近年来单颗粒冷冻电镜三维重构技术取得了长足的进展，已经成为结构生物学研究中至关重要和不可替代的研究手段，然而，在计算方面仍然存在很多挑战性的问题，制约着单颗粒冷冻电镜的进一步发展。特别是需要进行以下两方面的研究：

1. **快速准确的颗粒图像识别算法**

由于电镜使用了低剂量的电子辐射使得蛋白颗粒图像的信躁比非常低。要提高信躁比，得到高分辨率的分子三维模型，必须采集更多的电子显微颗粒图像。一般要获得原子分辨率的结构需要几十万张原始颗粒图像数据[18]。目前最可靠的颗粒图像挑选手段还是手工挑选。可以想象，如果手工挑选几十万张颗粒图像将是一件不可能的事情。颗粒图像识别算法一直是冷冻电镜三维重构的研究热点之一。《结构生物学杂志（*Journal of Structural Biology* ）》145 (2004)这一期集中总结了颗粒图像自动识别的方法，将其归纳为五类[10]：1. 基于模板（Template-based）的方法、2. 基于边沿检测（edge detection-based ）的方法、 3. 强度比较（Intensity comparison）方法、 4. 基于纹理（Texture-based）的方法、 5. 神经网络方法。虽然近年来在该领域的研究取得了重要的进展[19]，颗粒图像识别自动采集算法依

然是单颗粒三维重构的一个瓶颈，因此迫切需要发展快速准确的颗粒图像挑选算法。

2. 高性能计算

在冷冻电镜三维重构处理中每一张蛋白颗粒图像都需要通过计算确定颗粒的投影方向，然后利用中心截面定理和傅立叶变换获得蛋白颗粒的三维结构，且三维重构的模型还需要经过多次迭代优化，因此冷冻电镜三维重构极其耗时，通常需要10^6CPU 小时。例如 2008 年 2 月在《自然（*Nature*）》上发表的 ε15 噬菌体病毒的结构，就是基于 36,259 张颗粒图像，在普度（Purdue）大学的 Condor computing resource 上耗时几个月的时间（10^6CPU 小时）重构出来的，其结构分辨率为 4.5 埃[20]。现有的计算能力和计算方法已经根本无法对冷冻电镜试验数据进行及时的处理，严重制约了其在实际中的应用。因此利用高性能超级计算环境和计算技术开发快速准确的三维重构计算方法就显得尤为关键和迫切。

3 我们的工作

本节重点介绍我们课题组在冷冻电镜三维重构方面的一些研究工作。首先，介绍在冷冻电镜三维重构高性能计算方面的研究，其次介绍我们在蛋白颗粒识别方面的研究。

3.1 并行单颗粒重构软件 ParaEMAN

冷冻电镜单颗粒三维重构的处理涉及到多个计算模块，具有多样性的特点，因此一个单一的、全局的并行策略并不适合其并行化。针对 EMAN 处理流程的主要模块（投影、分类、平均、重构），我们分别提出了不同的并行策略，进行了并行单颗粒重构软件 ParaEMAN 在曙光 5000A 高性能计算系统的 1024 个核上的运行测试，得到了 508.5 倍的加速比。

ParaEMAN 设计实现中的核心问题是计算任务的动态调度：N 个颗粒被 ***classesbymra*** 程序分为 n 个类，其中第 i 个类中包含 k_i 个颗粒，之后由 ***classalign2*** 程序计算每一类的平均。如果按照常规的方式并行 ***classalign2***，需要将第 i 个类中的 k_i 个颗粒进一步划分为若干部分，分配给各进程或线程。该策略的实际结果并不理想，原因是各个类中的颗粒数量不均衡。对一个颗粒数量较少的类，每一个进程或线程得到的数据量将会很小，因而并行计算带来的性能提升将无法补偿多线程或多进程的开销，从而造成整体性能下降。针对这一问题，我们课题组提出了一种自适应动态调度策略(Self-Adaptive Dynamic Scheduling，SADS)有效实现了分类操作的并行化[21,22]。

SADS 将每个类的平均操作作为一个独立的任务，分配给不同的进程。与常规并行方式的每次并行处理一个分类操作不同，该策略一次考虑所有类的平均操作，根据各操作所需的处理时间，将相应的类分配给某个进程，同时使得各个进程的总处理时间尽可能相等。因此首要的问题就是估算各任务所需的处理时间。由于平均操作的时间复杂度为 $O(k_i)$，其中 k_i 代表第 i 个类中颗粒图像的个数。因此可用如下线性函数对任务的处理时间建模：

$$p_i = ak_i + b \quad i = 1,2,\cdots,m$$

其中 a 和 b 为待定系数，其值与系统配置有关。在实际研究中我们利用上一轮迭代中任务的实际执行时间来更新待定系数的值。假设在第 j 轮迭代中，待定系数的值为 $a(j)$ 和 $b(j)$，第 i 个类中颗粒图像个数为 $k_i(j)$，SADS 的框架如下：

1 置初值 $a^{(0)} = 1$，$b^{(0)} = 0$

2 for j=0 to ITER – 1 do

2.1 估算各任务的处理时间

$$p_i^{(j)} = a^{(j)} k_i^{(j)} + b^{(j)},\quad i = 1,2,\cdots,m$$

2.2 根据估算的任务处理时间 $p_i^{(j)}$ 分配任务给不同进程

2.3 各进程执行任务并记录实际运行时间 $t_i^{(j)}$

2.4 利用任务实际运行时间更新下一轮迭代待定系数的值：

$$a^{(j+1)} = \frac{m\sum_{i=1}^{m} k_i^{(j)} t_i^{(j)} - \sum_{i=1}^{m} k_i^{(j)} \sum_{i=1}^{m} t_i^{(j)}}{m\sum_{i=1}^{m} \left(k_i^{(j)}\right)^2 - (\sum_{i=1}^{m} k_i^{(j)})^2}$$

$$b^{(j+1)} = \frac{m\sum_{i=1}^{m} t_i^{(j)} \sum_{i=1}^{m} \left(k_i^{(j)}\right)^2 - \sum_{i=1}^{m} k_i^{(j)} \sum_{i=1}^{m} k_i^{(j)} t_i^{(j)}}{m\sum_{i=1}^{m} \left(k_i^{(j)}\right)^2 - (\sum_{i=1}^{m} k_i^{(j)})^2}$$

endfor

在以上算法框架中，2.2 步根据估算的任务执行时间将 m 个任务分配给 p 个进程，并使得进程的负载尽可能均衡。该问题可归结为整数规划：

Min t

$$\sum_{i=1}^{m} p_i x_{ij} \le t \quad j = 1,2,\cdots,m$$

$$\sum_{j=1}^{p} x_{ij} = 1 \quad i = 1,2,\cdots,m$$

$$x_{ij} \in \{0,1\} \quad i = 1,2,\cdots,m \quad j = 1,2,\cdots,p$$

这是一个具有 NP 难度的最优化调度问题。通过对该问题的适当简化，采用动态规划的方法求解，我们获得了较好的负载平衡。

3.2 冷冻电镜单颗粒图像识别方法

由于冷冻电镜颗粒图像的信噪比非常低，以及分子颗粒取向的随机性，给冷冻电镜分子颗粒挑选带来很大困难。针对这一问题，我们采用多种方法进行了多种颗粒图像识别算法的尝试，包括基于直方图信息熵、改进的 AdaBoost 算法[2]、贝叶斯分类、最小距离分类和相关性匹配等方法，有效降低了颗粒图像挑选的存伪比率（False Positive Rate，FPR）和弃真比率（False Negative Rate，FNR），主要的工作为：

- 根据颗粒区域与非颗粒区域的灰度直方图分布不同，同种颗粒区域之间灰度直方图分布较为相似的特点，我们提出了基于直方图信息熵的图像识别方法。信息熵采用以下计算方式：

$$V1 = \sum_{i=1}^{n} (f_1(i) \times \mathbf{log} f_1(i) - f_2(i) \times \mathbf{log}\, f_2(i))^2$$

$$V2 = \sum_{i=1}^{n} f_1(i) \times \mathbf{log}\left(f_1(i) / f_2(i)\right)$$

[2] Adaptive Boosting，一种机器学习算法。

其中，$f_1(i)$，$f_2(i)$ 分别代表模板与带识别的区域灰度分布。n 代表灰度级。V1 表示模板与带识别的区域灰度直方图信息熵的差值。V2 表示相对熵（或称 Kullback－Leibler 距离，可表示图像之间的差异）。

- 借鉴人脸识别中常用的 AdaBoost 算法，并结合冷冻电镜图像的特点，提出了利用分治原理优化 AdaBoost 的方法，提高了 AdaBoost 算法识别的精度。其思想是在 AdaBoost 算法的学习训练阶段，对整个样本集进行分治学习，并以子样本所占比重作为权重，对每个子样本生成的子强分类器进行组合[23]。

对每一个子样本训练一强分类器：

$$h_j(x)=\begin{cases}1, & f_j(x)>\theta_j\\ 0, & f_j(x)\le\theta_j\end{cases},\quad j=1,2,\cdots,T$$

$$H^n(x)=\begin{cases}1, & \text{如}\sum_{t=1}^{T}a_t h_t(x)\ge\frac{1}{2}\sum_{t=1}^{T}a_t\\ 0, & \text{其它}\end{cases},\qquad a_t=\log\frac{1-\varepsilon_t}{\varepsilon_t}$$

其中 $h_j(x)$ 表示弱分类器的值，θ_j 表示弱学习算法寻找出的阈值，$f_j(x)$ 表示特征值，x 表示一个 Haar 特征，ε_t 表示弱分类器的错误概率，n 表示第 n 个子样本的强分类器。

全样本空间的强分类器由子样本分类器的线性组合形成：

$$H(x)=\sum_{i=1}^{n}w_i H^i(x)$$

其中 $H(x)$ 为全样本空间分类器，w_i $(i=1,2,\cdots,n)$ 为每个子样本对应的权重。

- 借鉴贝叶斯分类器和最小距离分类器在特征分类方面的优势，提出了基于贝叶斯分类器和最小距离分类器的分类方法。

贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。这里用的是高斯模式类的贝叶斯分类器，即分类的模式服从高斯密度。

二维贝叶斯判别函数：

$$d_j=P(x/w_j)P(w_j)=\frac{1}{\sqrt{2\pi\sigma_j}}e^{\frac{-(x-m_j)^2}{2\sigma_j^2}}P(w_j),\ j=1,2,\cdots$$

其中 w_j 表示第 j 类模式，P 表示概率。

距离分类器利用分子颗粒在多维特征空间中呈现聚类分布的性质，在多维空间判断待识别的区域到聚类中心的距离。

$$D_j(x)=\|x-m_j\|,\ \text{其中}m_j=\frac{1}{N_j}\sum_{x\in w_j}x_j,\ j=1,2,\cdots,W$$

x 表示一未知的模式向量，N_j 是类 W_j 的模式向量数目，W 是模式类的数目。

我们在公共的冷冻电镜颗粒图像基准测试集[24]上就上述几种方法进行了测试，其挑选结果的存伪比率、弃真比率和识别率如表 1 所示。

表 1

算法	存伪比率	弃真比率	识别率
Adaboost	0.031	0.107	0.893
改进的 Adaboost	0.075	0.044	0.956
贝叶斯分类器	0.055	0.12	0.88
最小距离分类器	0.082	0.11	0.89
相关性匹配	0.0167	0.0647	0.9353

4 总结与未来工作

本文详细介绍了单颗粒冷冻电镜三维重构的发展和现状，分析了当前所面临的主要挑战性问题，着重介绍了我们在单颗粒冷冻电镜三维重构相关研究工作上的进展：提出了一种自适应动态调度策略（SADS），有效解决了冷冻电镜三维重构中的计算任务调度问题；并且在此基础上，开发完成了冷冻电镜三维重构的并行软件 ParaEMAN；实现了多种颗粒图像识别算法，有效降低了颗粒图像挑选的存伪率和弃真率。

在下一步的工作研究中，我们将主要在以下三个方面继续深入开展研究：（1）. 研究高性能的冷冻电镜三维重构算法，进一步完善 ParaEMAN；（2）. 研究快速准确的颗粒图像识别算法，进一步提高颗粒图像的识别率；（3）. 开展单颗粒冷冻电镜三维重构新算法的研究，例如基于球谐函数的三维重构算法。

参考文献

[1] Sali, R. Glaeser, T. Earnest, and W. Baumeister. From words to literature in structural proteomics, *Nature*, vol. 422(6928), pp. 216–225, 2003.

[2] J. Frank, *Three Dimensional Electron Microscopy of Macromolecular Assemblies*. London: Oxford Univ. Press, 2005.

[3] Xuekui Yu, Lei Jin, Z. Hong Zhou. 3.88Å structure of cytoplasmic polyhedrosis virus by cryo-electron microscopy. *Nature*, vol. 453(7193), pp.415-419, 2008

[4] Z. Hong Zhou. Towards atomic resolution structural determination by single-particle cryo-electron microscopy. Curr Opin Struct Biol. Vol.18(2), pp. 218-228, 2008.

[5] De Rosier, D., and Klug, A. Reconstruction of 3-dimensional structures from electron micrographs, *Nature*, vol. 217, pp.130–134 ,1968

[6] Taylor K, Glaeser RM. Electron diffraction of frozen, hydrated protein crystals. *Science*, vol.186, pp.1036–37, 1974.

[7] 王大能 ,陈勇 ,隋森芳.电子显微学在结构生物学研究中的新进展.电子显微学. Vol. 22(5), pp.449-456, 2003.

[8] J.-J. Fernandez, C.O.S.Sorzano, R. Marabini and J.-M.Carazo. Image processing and 3-D reconstruction in electron microscopy. *IEEE Signal Processing Magazine*. Vol. 23(3), pp. 84-94. 2006.

[9] Chiu Wah, et.al. Visualization of Biological Nano-Machines at Subnanometer Resolutions, *JEOL News*, Vol. 41(1), pp. 12-17, 2006.

[10] C.S. Potter, Y. Zhu, and B. Carragher, Eds., *J. Struct. Biol. (Special Issue on Automated Particle*

Selection for Cryo-Electron Microscopy), vol. 145, no. 1–2, pp. 1–180, 2004.

[11] J.A. Velázquez-Muriel, C.O.S. Sorzano, J.J. Fernández, and J.M. Carazo, "A method for estimating the CTF in electron microscopy based on ARMA models and parameter adjusting," *Ultramicrosc.*, vol. 96, no. 1, pp. 17–35, 2003.

[12] E.V. Orlova, P. Dube, J.R. Harris, E. Beckman, F. Zemlin, J. Markl, and M. van Heel, Structure of KLH1 at 15 Å resolution by electron cryomicroscopy and angular reconstitution, *J. Mol. Biol.*, vol. 271, no. 3, pp. 417–437, 1997.

[13] Liya Fan, Fa Zhang, Gongming Wang and Zhiyong Liu. A Framework to Refine Particle Clusters Produced by EMAN. *Bioinformatics*, vol. 25(12), pp. i276-280, 2009.

[14] A. Pascual-Montano, L.E. Donate, M. Valle, M. Bárcena, R. Pascual-Marqui, and J.M. Carazo. A novel neural network technique for analysis and classification of EM single particle images. *J. Struct. Biol.*, vol. 133(2–3), pp. 233–245, 2001.

[15] S.H.W. Scheres, M. Valle, R. Núñez, C.O.S. Sorzano, R. Marabini, G.T. Herman, and J.M. Carazo, Maximum-likelihood multi-reference refinement for electron microscopy images. *J. Mol. Biol.*, vol. 348(1), pp. 139–149, 2005.

[16] Steven J. Ludtke, Philip R. Baldwin, and Wah Chiu. EMAN: Semiautomated Software for High-Resolution Single-Particle Reconstructions, *J. Struct. Biol.*, vol. 128(1), pp. 82-97, 1999.

[17] 樊莉亚，张法，王功明，刘志勇。单颗粒重构软件 EMAN 算法分析与高效并行实现，计算机研究与发展（在审）

[18] Richard Henderson. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly Reviews of Biophysics*, Vol. 28(2), pp. 171-193, 1995.

[19] Stagg, S.M., Pulokas, J., Fellmann, D., Cheng, A., Quispe, J.D., Mallick, S.P., Avila, R.M., Carragher, B., Potter, C.S., 2006. Automated cryoEM data acquisition and analysis of 284,742 particles of GroEL. *Nature* Vol.439, pp.234–238, 2006.

[20] Jiang W, Baker ML, Jakana J, Weigele P, King J, Chiu W. Backbone structure of the infectious epsilon15 virus capsid revealed by electron cryomicroscopy. *Nature*, Vol. 451, pp. 1130-1134, 2008.

[21] Liya Fan, Fa Zhang, Gongming Wang, Bo Yuan, Zhiyong Liu. A Self-Adaptive Greedy Scheduling Scheme for a Multi-Objective Optimization on Identical Parallel Machines. *Studies in Computational Intelligence*, Vol.209, pp.43-55, 2009.

[22] Liya Fan, Fa Zhang, Gongming Wang, Zhiyong Liu. An Effective Scheduling Algorithm for Linear Makespan Minimization on Unrelated Parallel Machines. *The 16th annual IEEE International Conference on High Performance Computing (HiPC 2009)*, Accepted.

[23] 邵书伟，张法，邱显杰，王兆其，刘金刚，孙飞。基于旋转和平移不变性的 EM 分子颗粒挑选算法，（in prepare）

[24] Zhu, Y., Carragher, B., Mouche, F., Potter, C.S. Automatic particle detection through efficient Hough transforms. *IEEE Trans. Med. Imaging*, Vol.22, pp. 1053-1062, 2003.

作者简介：

樊莉亚： 中国科学院计算技术研究所前瞻研究中心研究生
邵书伟： 中国科学院计算技术研究所前瞻研究中心研究生
王功明： 中国科学院计算技术研究所前瞻研究中心研究生
万晓华： 中国科学院计算技术研究所前瞻研究中心研究生
储　琪： 中国科学院计算技术研究所前瞻研究中心研究生
陈　翔： 博士，中国科学院计算技术研究所前瞻研究中心助理研究员
张　法： 博士，中国科学院计算技术研究所前瞻研究中心副研究员，zf@ncic.ac.cn